Magyar Tudomány 179(2018)7, 955–963
DOI: 10.1556/2065.179.2018.7.4

HIBRID NYELVTECHNOLÓGIÁK

HYBRID HUMAN LANGUAGE TECHNOLOGIES

Ács Judit[1], Borbély Gábor[2], Makrai Márton[3], Nemeskey Dávid[4], Recski Gábor[5], Kornai András[6]

[1] tanársegéd, Budapesti Műszaki és Gazdaságtudományi Egyetem Automatizálási és Alkalmazott Informatikai Tanszék
[2] tanársegéd, Budapesti Műszaki és Gazdaságtudományi Egyetem Algebra Tanszék
[3] tudományos segédmunkatárs, MTA Nyelvtudományi Intézet
[4] informatikus mérnök, MTA SZTAKI Nyelvtechnológiai Kutatócsoport
[5] PhD, tanársegéd, Budapesti Műszaki és Gazdaságtudományi Egyetem Automatizálási és Alkalmazott Informatikai Tanszék
[6] az MTA doktora, az MTA SZTAKI Nyelvtechnológiai Kutatócsoport vezetője, tudományos tanácsadó, a Budapesti Műszaki és Gazdaságtudományi Egyetem Algebra Tanszék professzora, kornai@sztaki.hu

ÖSSZEFOGLALÁS

Az elmúlt harminc év nyelvészetét a „racionalista” (szabályalapú, szimbólumkezelő) és az „empirista” (statisztikai alapú, gépi tanulásos) nyelvészeti modellek harca jellemezte. Míg a nyolcvanas években még egyértelműen a racionalista paradigma volt az uralkodó, mára ez, különösen az utolsó néhány év mélytanulásos forradalmának köszönhetően megfordult, és egyértelműen az empirista paradigma lett a domináns. Az MTA SZTAKI nyelvtechnológiai csoportja elsősorban a hibridizáció kérdéseivel foglalkozik, azzal, hogy miképp találhatjuk meg a diszkrét, szimbolikus struktúrát a folytonos, zajos adatokban, illetve hogyan tudjuk a struktúráról való ismereteinket hatékonyabb algoritmusok építésében kamatoztatni.

ABSTRACT

In the last thirty years linguistics was characterized by a debate between the “rationalist” (rule-based, symbol-manipulating) and the “empiricist” (statistics-based, machine learning) camps. Back in the 1980s clearly the rationalist paradigm had the upper hand, but by now the situation is reversed, and thanks to the deep learning revolution of the past few years, today the empiricist paradigm dominates. The human language technology group at MTA SZTAKI focuses on issues of hybridization, in particular on finding the discrete symbolic structure in the continuous (and noisy) data, and on leveraging our knowledge of structure in building more efficient algorithms.

Kulcsszavak: nyelvtechnológia, gépi tanulás, mélytanulás, hibrid rendszerek

Keywords: human language technology, machine learning, deep learning, hybrid systems

BEVEZETÉS

Az MTA SZTAKI Nyelvtechnológiai (Human Language Technology, HLT) Kutatócsoportjának előzményei az origo.hu (Origo) és a Northern Light Technologies (NLT) közti együttműködés időszakára nyúlnak vissza. Ma az Origo csupán egy a számtalan webes portál közül, de 2002-ben, amikor az együttműködés az addig használt AltaVista (AV) keresőtechnológia tarthatatlansága miatt szükségessé vált, az Origo még úgy uralkodott a magyar weben, mint a 19. században Britannia a habok felett: látogatottsága nagyobb volt, mint az őt követő két legnagyobb portálé együttvéve. Az NLT, melynek akkoriban Kornai András volt a tudományos vezetője, 1999-ben nőtt nagyobbra, mint az AltaVista (Yahoo), és kettejük versenyében (melyet végül a nevető harmadik, a Google nyert meg) már tetten érhető volt az a szemléletbeli különbség a racionalista és az empirista megközelítések közt, amelyet pár évvel korábban már igen markánsan jelzett Judith L. Klavans és Philip Resnik (1996).

Míg az AV (web yahoo-knak nevezett) szerkesztők százait foglalkoztatta, akik szabályalapon kézzel sorolták be a weblapokat eleinte néhány tucat, később több ezer, hierarchikusan elrendezett tartalmi kategóriába, addig az NLT statisztikai módszerekkel alakította ki az egyes kategóriák modelljeit, és mivel a besorolás teljesen automatikus volt, nem volt szükség a szerkesztői gárdának a web robbanásszerű növekedését követő bővítésére (mely végső soron a Yahoo/AV vesztét is okozta). A magyar tematikus hierarchia úttörője Ungváry Rudolf (Országos Széchenyi Könyvtár) volt, az Origóban használt rendszert az ő munkáját továbbfejlesztve dolgozta ki Kárpáti András és Halácsy Péter (ma a Pécsi Tudományegyetem, illetve a Prezi, akkoriban az Axelero, a mai T-Online munkatársai). Az NLT az általuk készített katalógus mint tanulóadat alapján építette fel a saját modelljeit gépi tanulásos módszerekkel [Kornai, 2003 EACL][1]. Mint ismeretes, a gépi tanulást (machine learning) máig a címkézett adatokon alapuló ún. felügyelt tanulás (supervised learning) dominálja. A nyers adatokon alapuló felügyeletlen (unsupervised) tanulás nagy erőkkel kutatott terület, ahol komoly eredményekről csak az utóbbi tíz évben beszélhetünk (Erhan et al., 2010), és az igazi áttörés, a felügyeletlen struktúratanulás, még várat magára.

Ebben a cikkben a kutatásoknak a Műegyetemen otthont adó Média Oktató és Kutató Központtal (MOKK) nem tudunk annak jelentőségéhez mérten foglalkozni, bár kétségkívül ez volt a számítógépes társadalomtudomány első multidiszciplináris műhelye hazánkban, ahol a számítógépes nyelvészet csupán egy volt a digitális szerzői joggal, kulturális termeléssel, a digitális térrel és annak

[1] A munkacsoport azon cikkei, melyek hivatkozásai [...] közt szerepelnek, elérhetőek a HLT honlapján: https://hlt.bme.hu/en/publications, a (...) közti hivatkozásokat lásd a cikk végén lévő irodalomjegyzékben.

szociológiájával, a formális cselekvéselmélettel, az új médiával, *peer to peer* hálózatokkal stb. foglalkozó kutatások közül. Reméljük, hogy a nemrég a Preziben Babarczy Eszter, Bodó Balázs, Csigó Péter, György Péter, Halácsy Péter, Kacsuk Zoltán, Szakadát István, Varga Dániel és Vályi Gábor részvételével megrendezett MOKKtóber találkozó anyagai megteremtik az alapot e műhely történetének és máig érezhető hatásának alaposabb feltárásához.

Az akkori nyelvtechnológiai munkák közül megemlítjük az első magyar szabadon letölthető korpuszt (WebKorpusz), az első párhuzamos magyar–angol korpuszt (hunglish.hu), és a Hun* eszközláncot, melyek az első nyílt forráskódú (open source) magyar nyelvi szoftverek közt voltak. Ebbe az eszközláncba épült be az eredetileg Németh László által külön fejlesztett HunSpell helyesírás-el lenőrző is, mely azóta is a szabad világ vezető helyesírás-ellenőrzője (ugyanaz a szoftver-keret több mint száz nyelvhez lett adattal feltöltve, és ma megtalálható a Thunderbird, FireFox, LibreOffice sok millió példányában); a Simon Eszter által épített HunNER névelem-felismerő [Simon, 2013; Nemeskey–Simon, 2012], és még sok más eszköz, melyekről az alábbiakban részletesen lesz szó. A mokk.bme.hu és a nyelvtechnológiai vonalon ezt továbbvivő hlt.bme.hu máig a nyílt forráskódú nyelvtechnológia egyik vezető képviselője, azzal a fontos különbséggel, hogy az elmúlt másfél-két évtizedben megfordult a széljárás, és az egykor ignorált, majd kinevetett, majd ellenségnek tekintett nyílt forráskódú megközelítés mára uralkodóvá vált.

KÉTFÉLE SZEMLÉLET

Tudományszociológiai szemszögből a racionalista és az empirista kutatási modellek közti különbség lényege a felülről vezérelt (top down) és az alulról kiinduló (bottom up) keresési stratégia. Előbbi klasszikus példája a Manhattan Project, amely a fizikusok elismert vezetőjének, Albert Einsteinnek az elnökhöz intézett levele alapján indult be: legfelül pár tucat elméleti fizikus, alattuk több száz mérnök és kísérleti fizikus, akik alatt munkások ezrei dolgoztak. A nyelvészetnek is megvolt a maga elismert vezetője, Noam Chomsky, aki nagyon is határozott irányú kutatásokat kezdeményezett. Annak az egyszerű, de előtte kevéssé hangsúlyozott ténynek az alapján, hogy a kisgyermekek viszonylag gyorsan, néhány év alatt lényegében tökéletesen megtanulják anyanyelvüket (és bármely nyelvi környezetbe helyezzük a csecsemőt, az ottani nyelvet képes ilyen szinten megtanulni), arra a következtetésre jutott, hogy ennek a tanulási képességnek kizárólag az lehet a magyarázata, hogy a gyermek fejében a tudásanyag egy nagy része, az univerzális grammatika, már örökletesen ott van.

Bár kezdettől voltak ennek az elméletnek komoly ellenzői, például Jean Piaget (Chomskyval való vitájának hiteles összefoglalóját adja Piattelli-Palmarini

et al., 1980), nyugodtan elmondhatjuk, hogy a fentebb idézett nagy hatású publikációktól kezdve a modern nyelvészeti kutatások fővonalát a 20. században Chomsky jelölte ki [Kornai, 2010 HRP], és nem kevesek számára máig az ő felfogása szolgál iránytűként. De a *Zeitgeist* megváltozott, a bölcs vezetők kora lejárt, és ami a legfontosabb: a predikciók nehezen megfoghatónak bizonyultak, specifikus nyelvtani struktúrákat/géneket nem sikerült azonosítani a szótan és mondattan területén. A kudarc annál is fájóbb volt, mert a hangtanban frappáns csecsemőkísérletek sora (összefoglalásukat lásd Werker–Tees, 1984) nyilvánvalóvá tette, hogy Chomskynak igaza van: az egyes nyelvek hangtanának kisgyermekkori elsajátítása nem magyarázható univerzális fonetika tételezése nélkül.

Ez a megváltozott Zeitgeist tette lehetővé, hogy a terméketlennek bizonyult elméleti megfontolásokból nagyrészt kiábrándult nyelvészek egyre komolyabban vegyék a lentről, a kutatás lövészárkaiból érkező empirikus anyagot. Egyre nagyobb és nagyobb egy- és többnyelvű korpuszt lehetett számítógépes elemzés alá vetni. A bevezetőben már érintettük azokat a korpuszfejlesztési munkálatokat, melyeket a HLT-csoport végzett. Ezek jelentősége nem pusztán abban áll, hogy az addigi nagyon komoly és szakmailag jól megalapozott korpuszokat, mint például a Magyar Nemzeti Szövegtár akkori változata (Váradi, 2002) vagy az elemzett (és ezért természetesen jóval kisebb) Szeged Korpusz (Vincze et al., 2014) nyíltabbá, jobban elérhetővé tette (ezt inkább a megváltozott Zeitgeistnek, mint a Webkorpusz és a Hunglish megjelenésének tudjuk be), hanem abban, hogy elődeiknél lényegesen nagyobbak voltak.

A modern számítógépes elemzés legfontosabb alapanyagát a milliárdszavas (gigaword) korpuszok adják. Azok az elemzési technikák, melyek ma a kutatást uralják, kisebb anyagokon egyszerűen nem működnek jól. A legfontosabb elméleti újítás, mely az utóbbi öt-tíz évben áttörést hozott számos olyan területen, mint a képek és nyelvi leírásuk (caption) közti szemantikai kapcsolat gépi tanulása (Karpathy et al., 2014), a szóvektorok (embedding) bevezetése volt. Minden szóhoz egy véges (általában pár száz) dimenziós vektort rendelünk úgy, hogy a hasonló kontextusokban szereplő szavak vektorai egymáshoz hasonlóak (euklideszi térben közeliek) legyenek. Az első áttörést Ronan Collobert és szerzőtársai (2011) hozták el, akik egyszerre, ugyanazon vektorok felhasználásával, tudtak javítani több olyan klasszikus feladat addigi legjobb eredményén, mint a szófaj szerinti címkézés (part of speech tagging), a névelem-felismerés (named entity recognition), a sekély mondattani elemzés (tehát a mondatok pszichológiailag releváns darabokra, például főnévi csoportokra bontása [chunking]) és a szemantikai szerep felismerése (semantic role labeling). A kulcsmomentum itt az, hogy Colloberték nem egy új feladatot oldottak meg az új reprezentációval, hanem már régről ismert, nehéz, kutatók százai által vizsgált feladatokra (melyek többségével csoportunk is foglalkozott, például a HunTag szekvenciális címkéző [Halácsy

et al., 2006 LREC] vagy a sekély mondattani elemzés, mely máig aktív témánk [Recski, 2014]) értek el az eddigieknél jobb eredményeket.

A szemantika területén, ahol régen, évtizedekig előre hatóan a vezető kutatók, Chomsky és Richard Montague jelölték ki a kutatás fő irányát, ma a kutatók többsége egy olyan jelenséggel foglalkozik, amelyet egy brünni műegyetemista, Tomas Mikolov fedezett fel: a szóvektorok lineáris struktúrát mutatnak, például $v(king) - v(man) + v(woman) \approx v(queen)$ (Mikolov et al., 2013b). Csoportunk a vektoros szófordítás (lineáris fordítás, lásd Mikolov et al., 2013a) módszerét alkalmazta közép-európai nyelvekre [Makrai et al., 2013], olyan ritkábban vizsgált lexikai relációk felé általánosítottuk az analógia vektoralgebrai megfogalmazását, mint a jó-rossz (peace-war, pleasure-pain) vagy a föntlent (tall-short, rise-fall), Makrai Márton [2014 MSZNY] pedig oksági párok (például sérül-fáj) geometriáját elemezte. Új módszereket vezettünk be többjelentésű beágyazások (multi-sense embeddings) szemantikai felbontóképességének mérésére [Borbély, 2016 RepEval]. Ezekben a reprezentációkban egy-egy szóalakhoz több vektor is tartozhat, melyek elvileg a szó különböző jelentéseinek felelnek meg. A gyakorlatban azonban a jelentésvektorok között nem mindig figyelhető meg fogalmi különbség, egy-egy általánosabb vektor több jelentést is lefed, és fölösleges vektorok is lehetnek, melyek a modellnek egy alkalmazásban való hasznosságát ronthatják.

Utólag természetesen megtalálható a szóvektorok használatának elméleti megalapozása: a kontextus nyilvánvalóan fontos, és a gondolat, hogy egy szó jelentését a használati kontextuson keresztül érdemes megragadni, kétségkívül jelen van már a nagy brit strukturalista, John Rupert Firth munkáiban is, aki azt írta, „a word is characterized by the company it keeps" (a szavakat a társaságuk jellemzi). Ugyanakkor világosan kell látni, hogy Firth (akinek a prozódiára vonatkozó felfogása is újra életre kelt a modern fonológiában, lásd Goldsmith, 1990) éppen ahhoz az iskolához tartozik, amely ellen Chomsky egész életében harcolt. A nagy tömegű adat viszont minden területen a strukturalistákat, nem pedig az elsősorban szellemes anekdotikus példákra és nyelvi intuícióra alapozott chomskyánus megközelítést látszik igazolni.

HIBRID MODELLEK

A fentiek után talán meglepően hangzik, de korunkban az egész nyelvészet Chomsky programját követi két alapvető tekintetben is. Az egyik a már Chomsky (1965) által középpontba állított magyarázó adekvátság (explanatory adequacy) elve, mely szerint a nyelvelmélet nem állhat meg a tények leírásánál, hanem arra is magyarázatot kell adnia, hogy a kisgyermek hogyan sajátítja el a nyelvet, a másik az univerzálék (minden nyelvre egyaránt igaz állítások) keresése, melynek

Joseph Greenberg (1963) után szintén Chomsky fentebb vázolt programja adott új lendületet.

A legfontosabb különbség nem a generatív felfogásban, hanem az univerzális metaelméletet konkrétan realizáló nyelvtanok technikai apparátusában van. A szintaxis területén ez azt jelenti, hogy a környezetfüggetlen mélyszerkezeten és az ezt mozgató faátalakításokon alapuló transzformációs grammatika helyét átvette egy másik, szintén a strukturalista korszakból átvett formalizmus, a függőségi grammatika (Tesniére, 1959). Ebben az elméleti keretben ma már ötven nyelvhez találunk komoly, elemzett fabank (treebank) korpuszokat, jelenleg hetvenet, de számuk egyre nő (URL2). Ezek egységesített (univerzális) szófaj- és függőségtipológián alapulnak, és ezzel nagyban elősegítik a minden emberi nyelvre kiterjedő univerzálékutatást. Az empirikus alapok kiterjesztésére mindig is megvolt a szándék: már Greenberg is harminc nyelvvel dolgozott, de nyersanyagául nyelvtani leírások, nem pedig a direkt empirikus adatok szolgáltak. Tekintve, hogy mintegy hat-hétezer emberi nyelvről tudunk (bár ezekből gigaword korpuszra és fabankra a digitális nyelvhalál miatt legfeljebb háromszáznál számíthatunk [lásd Kornai, 2013 PLoS], az univerzális grammatika kutatása még sok évtizedre fog programot adni a nyelvészetnek.

Az új technikai apparátusra való áttérés egyébként a fonológiában is végbement, ahol a környezetfüggő, szekvenciális szabályrendszereket egy véges automatákkal megfogalmazható elmélet, az optimalitás elmélete váltotta fel (Prince–Smolensky, 1993; Karttunen, 1998). A technika megváltozása jelentős átalakulást hozott a szemantikában is, ahol a logikai formán (első- vagy magasabb rendű predikátumkalkuluson) alapuló reprezentációkat egy egyszerűbb, a függőségi fákkal egyenértékű függvényargumentum-szerkezet váltotta fel. Ezt tekinthetjük az ún. generatív szemantikához (Huck–Goldsmith, 1995) való visszatérésnek, de valójában sokkal régebbre, egészen az első formalizált nyelvtanig, Pāṇini *Aṣṭādhyāyī*-jáig (i. e. 450 körül) megy vissza.

Ebben a szellemben dolgoztuk ki 2009 és 2012 között a 4lang formalizmust [Kornai, 2010 MoL, 2012 LNCS; Kornai–Kracht, 2015; Kornai megj. alatt], mely a természetes nyelvi jelentést fogalmak irányított gráfjaként reprezentálja. Megalkottuk a text_to 4lang szoftvert [Recski, 2016 LREC], mely nyers angol és magyar szövegekhez automatikusan rendel ilyen reprezentációkat; ezeket sikerrel alkalmaztuk lexikális ontológiák építéséhez [Recski, 2016 LREC], és a fentebb Mikolov kapcsán már említett analógiás feladatok megoldásában [Recski et al., 2016 RepLearn]. Megemlítjük néhány a 4lang jelentés-reprezentációs rendszerhez [Kornai et al., 4th JCLCS] kapcsolódó kutatásunkat: az igei szerepek vizsgálata [Makrai, 2014 MSZNY], a definiáló szókincs [Kornai et al., 2015 MOL] és az aktivációterjedés [Nemeskey et al., 2013] kapcsán.

A magyarázó adekvátság tekintetében is ugyanez a folyamat játszódott le: az eszme győzedelmeskedett, de a technikai apparátus gyökeresen szembe megy a

Chomsky és Lasnik (1993) által javasolttal. Kicsi, néhány tucat diszkrét (bináris) paraméter beállításán alapuló döntési fák helyett nagy, sok százezer (gyakran sok millió) folytonos paraméter gradiens-módszerrel való tanulása vált uralkodóvá. Az ilyen sokparaméteres rendszerek tanulása a beszéd- és írásfelismerés terén indult be az ún. Rejtett Markov Modellek (Hidden Markov Model, HMM) felhasználásával: itt kapott először fontos szerepet a valószínűségi nyelvmodellezés (language modeling). Csoportunk mind a hagyományos (szó-n-eseken alapuló, n-gram), mind a mélytanulásban elterjedt rekurrens neurális háló alapú modelleket kutatja. Foglalkozunk a terület mind általános, mind a magyar nyelvre specifikus problémáival is [Nemeskey, 2017 MSZNY]. A természetes nyelvi mondatok hosszára valószínűségi, generatív modellt alkottunk, ami magyarázni tudja a mondatok empirikusan mérhető hosszeloszlását.

A magyar nyelv agglutinatív voltából fakadóan a szavak sok felszíni formában lehetnek jelen, ami az angol nyelvben jól működő szóalapú módszereknek komoly kihívást jelent. Vizsgálataink egyik fókusza annak megállapítása, hogy morfológiai eszközök mennyiben tudják ezt a problémát enyhíteni. OTKA-pályázat keretében vizsgáljuk a szavak és morfémák (legkisebb önálló jelentéssel rendelkező nyelvi egységek, például tárgyrag) neurális hálózatokkal történő azonosítását. A morfológiai elemzés számos nyelvtechnológiai feladat elengedhetetlen része, amit hagyományosan nyelvészek hosszas munkájával összeállított szabályok segítségével végeznek, azonban ezek a szabályok csak a világ nyelveinek töredékéhez állnak rendelkezésre. Kutatásunk célja olyan módszerek kidolgozása, amelyek pusztán nyers szövegből képesek ezeket a szabályokat felismerni. Bár ez a rendszer még nincs kész, előmunkálatai közül említést érdemelnek az automatikus szótárépítéssel [Ács et al., 2013, 2014] és ékezet-visszaállítással [Ács–Halmi, 2016] foglalkozó rendszereink.

Foglalkozunk a szóvektorok általánosításaival mátrix- és projektívtér-modellekre. A szokásos szó-vektoralapú beágyazások szisztematikus hibája (Pennington et al., 2014), hogy antonima-párok hasonló vektorokkal reprezentálódnak, például *good* ≈ *bad*. Ennek egy megoldását kínálja a projektív tér, ahol egy gömbön az antipodális pontok azonosítva vannak. Egy erre épülő célfüggvénnyel sikerült javítanunk a vektorbeágyazások által elért eredményt a Simlex999-adaton (Hill et al., 2014). A mátrixbeágyazások esetében egy szóhoz nem egy vektort, hanem egy mátrixot rendelünk. Ezzel egy nem-kommutatív általánosítását adjuk a szó-vektoroknak, melyek alkalmasak nyelvmodellezésre és speciális véges automaták tanítására is. A hibrid modellek diszkrét komponensei a mátrixmodellek, illetve az ezekkel szoros formai kapcsolatban álló véges automaták, melyek tanítása súlyozott nyelveken [Kornai et al., 2013] a szimbolikus és a probabilisztikus modellezésnek az eddigieknél mélyebb hibridizációját készíti elő.

ÖSSZEFOGLALÁS

A racionalista és az empirista megközelítések nem kizárják, hanem támogatják egymást. A modern gépi tanulás alapvető sikerkritériumai messze túlmennek a leíró adekvátságon (descriptive adequacy). A terület egyik legsikeresebb kutatócsoportja, Yoshua Bengio, Aaron Courville és Pascal Vincent (2013) külön kiemeli, hogy „In good high-level representations, the factors are related to each other through simple, typically linear dependencies" (a jól működő magas szintű reprezentációkban a tényezők egyszerű, tipikusan lineáris kapcsolatban állnak). Ez alól, úgy tűnik, a nyelvtan sem kivétel: a sikeres modellek mögött egyszerű lineáris nyelvtanokat (véges automatákat, véges transzducereket), illetve ezek olyan egyszerű általánosításait találjuk, mint a rejtett Markov-modellek vagy az Eilenberg-gépek (Eilenberg, 1974). A jövő útja, úgy véljük, az ilyenek automatikus tanulása, és ehhez, úgy tűnik, nincs semmilyen speciális, az ember általános kognitív képességein túlmutató eszközre szükség.

Azt gondoljuk, hogy a nyelvészeti vizsgálatok a számítógépes társadalomtudományok más területei számára is szolgálhatnak ilyen általános tanulságokkal, hiszen ezekben is az egyik fő cél a mögöttes struktúra feltárása, és ezekben is egyre inkább elérhetővé válik az a hatalmas tömegű adat, amelynek alapján e struktúra algoritmikus módszerekkel megragadható.

IRODALOM

Bengio, Y. – Courville, A. – Vincent, P. (2013): Representation Learning: A Review and New Perspectives. *IEEE Transactions* on Pattern Analysis and Machine Intelligence, 35, 8, 1798–1828. https://arxiv.org/pdf/1206.5538.pdf

Chomsky, N. (1965): *Aspects of the Theory of Syntax*. MIT Press, https://faculty.georgetown.edu/irvinem/theory/Chomsky-Aspects-excerpt.pdf

Chomsky, N. – Lasnik, H. (1993): Principles and Parameters Theory. *Syntax: An International Handbook of Contemporary Research*. (ed. Jacobs, J.) 1. Berlin: de Gruyter, 505–569.

Collobert, R. et al. (2011): Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research* (JMLR), 12, 2493–2537. http://www.jmlr.org/papers/volume12/collobert11a/collobert11a.pdf

Eilenberg, S. (1974): *Automata, Languages, and Machines*. Orlando, FL: Academic Press

Erhan, D. et al. (2010): Why Does Unsupervised Pre-training Help Deep Learning? *Journal of Machine Learning Research,* 11, 625–660. http://www.jmlr.org/papers/volume11/erhan10a/erhan10a.pdf

Goldsmith, J. A. (1990): *Autosegmental and Metrical Phonology.* Cambridge, MA: Blackwell

Greenberg, Joseph H. (1963): Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. *Universals of Human Language.* (ed. Greenberg, J. H.) MIT Press, 73–113. http://pkdas.in/JNU/typo/lu.pdf

Hill, F. – Reichart, R. – Korhonen, A. (2014): Simlex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *Computational Linguistics,* 41, 4, 665–695. https://arxiv.org/pdf/1408.3456.pdf

Huck, G. J. – Goldsmith, J. A. (1995): *Ideology and Linguistics Theory: Noam Chomsky and the Deep Structure Debates*. London: Routledge

Karpathy, A. – Armand, J. – Fei Fei, L. (2014): Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. *Advances in Neural Information Processing Systems, 27.* (ed. Ghahramani, Z. et al.) Curran Associates, Inc., 1889–1897. https://cs.stanford.edu/people/karpathy/nips2014.pdf

Karttunen, L. (1998): The Proper Treatment of Optimality in Computational Phonology: Plenary Talk. *Proceedings of the International Workshop on Finite State Methods in Natural Language Processing*. Association for Computational Linguistics, 1–12. https://arxiv.org/pdf/cmp-lg/9804002.pdf

Klavans, J. L. – Resnik, P. (eds.) (1996): *The Balancing Act – Combining Symbolic and Statistical Approaches to Language*. MIT Press

Mikolov, T. – Le, Q. V. – Sutskever, I. (2013a): Exploiting Similarities among Languages for Machine Translation. arXiv:1309.4168. https://arxiv.org/pdf/1309.4168.pdf

Mikolov, T. – Yih, W. – Zweig, G. (2013b): Linguistic Regularities in Continuous Space Word Representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT 2013). Atlanta, Georgia: Association for Computational Linguistics, 746–751. https://www.aclweb.org/anthology/N13-1090

Pennington, J. – Socher, R. – Manning, C. (2014): GloVe: Global Vectors for Word Representation. *Conference on Empirical Methods in Natural Language Processing (*EMNLP 2014). 1532–1543. https://www.aclweb.org/anthology/D14-1162

Piattelli-Palmarini, M. – Piaget, J. – Chomsky, N. (1980): *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Routledge

Prince, A. S. – Smolensky, P. (1993): *Optimality Theory: Constraint Interaction in Generative Grammar*. (*Rutgers University Center for Cognitive Science Technical Report* 2.) Piscataway, NJ: Rutgers University DOI:10.1002/9780470759400

Tesniére, L. (1959): Élements de syntaxe structurale. Paris: Klincksieck, https://archive.org/details/LucienTesniereElementsDeSyntaxeStructurale

Váradi T. (2002): The Hungarian National Corpus. *Proceedings of the Third International Conference on Language Resources and Evaluation*, 385–389. https://www.researchgate.net/publication/228608174_The_Hungarian_National_Corpus

Vincze V. et al. (2014): Szeged Corpus 2.5: Morphological Modifications in a Manually POS-tagged Hungarian Corpus. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* (eds). Nicoletta Calzolari (Conference Chair) et al. Reykjavik, Iceland: European Language Resources Association (ELRA), 1074–1078. http://publicatio.bibl.u-szeged.hu/4736/1/szk.pdf

Werker, J. F. – Tees, R. C. (1984): Cross-language Speech Perception: Evidence for Perceptual Reorganization during the First Year of Life. *Infant Behavior and Development*, 7, 49–63. DOI: 10.1016/S0163-6383(84)80022-3, https://bit.ly/2IJwyt8

URL1: hlt.bme.hu

URL2: universaldependencies.org